Повышение эффективности контроля знаний студентов на основании анализа последовательности решения тестовых заданий

Щербина Дмитрий Николаевич
доцент, к.б.н., Южный федеральный университет,
пр. Стачки, 192/1, г. Ростов-на-Дону, 344090, (918)5658761
dnsherbina@sfedu.ru

Аннотация

Метод повышения точности тестирования знаний за счет выделения стратегий выполнения тестовых заданий апробирован на студентах-биологах, проходивших тестирование по физиологии. Информация о стратегиях прохождения теста позволяет преподавателю точнее оценить уровень подготовки студента, пресечь завышение оценки из-за списывания и случайного угадывания. Повышение эффективности контроля знаний достигается за счет сокращения количества тестовых заданий, необходимых для получения оценки с той же надежностью.

The method of improving the accuracy of knowledge testing through the detection of test-taking strategy was tested on students of biology department. The additional information about strategies allows the teacher to more accurately assess the level of student preparation, to prevent overestimation due to cheating and random guessing. Improving the efficiency of the control of knowledge is achieved by reducing the number of test items required to obtain an estimate with the same reliability.

Ключевые слова

компьютерное тестирование, оценка знаний, moodle;
computer-based testing, knowledge tests, moodle.

Введение

Актуальная проблема организации образовательного процесса – обеспечение мониторинга текущего развития компетенций. Мониторинг заключается в регулярном тестировании знаний и навыков. Для надежной оценки нужны тесты из многих тестовых заданий [1]. Регулярные тесты, состоящие из многих заданий, требуют много времени. Задача повышения эффективности контроля знаний сводится таким образом к сокращению длительности тестирования без потери качества оценки.

Поскольку текущие знания студента по данной теме K не поддаются прямому измерению, то приближенную оценку знаний $\hat{K}$ получают усреднением n оценок за знание отдельных фактов, относящихся к данной теме. Знание отдельного факта k_i оценивается путем предъявления студенту задания z_i, где $i = 1 \ldots n$. Как правило оценка k_i бинарная: $k_i = 1$, если задание решено правильно; $k_i = 0$, если задание решено неправильно (о других моделях оценивания см. [2]). Усредненная оценка $\hat{K}$ представляет собой апостериорную вероятность того, что студент обладает знаниями по теме, исходя из биномиального распределения n полученных k.

Задача тестирования конкретного студента на знание отдельного факта k_i сводится к последовательному выполнению трех шагов:

1) генерация вопроса, для ответа на который требуется знание данного факта;

2) предъявление вопроса студенту;

3) получение и оценка ответа.

Эта процедура повторяется n раз до достижения стабильной оценки.

С точки зрения статистики, ответы на отдельные тестовые задания являются экземплярами из потенциально бесконечной генеральной совокупности всевозможных заданий данному индивиду по данной теме. Ограниченное число заданий представляет собой выборку, по которой мы пытаемся оценить средний уровень знаний по данной теме. Увеличивая размер выборки, мы снижаем ошибку среднего, убывающую с корнем квадратным от числа наблюдений, до пренебрежимо маленького размера.

Другой способ снижения ошибки среднего – за счет снижения вариабельности оценки (числитель в формуле стандартной ошибки). Снижения вариабельности оценки $\tilde{K}$ относительно K добиваются методами адаптивного тестирования, которое позволяет подбирать задания, совпадающие по сложности с текущим уровнем знаний [3, 4]. Признавая важность и большой потенциал применения адаптивного тестирования, в данной работе мы для простоты принимаем допущение, что все предъявляемые задания адекватны минимально требуемому уровню знаний студента, то есть их сложность нормально распределена вокруг этого уровня.

Повышение эффективности тестирования знаний сводится к уменьшению времени тестирования без потери точности суммарной оценки. Это возможно двумя путями:

1) сокращение времени на тестирование знания отдельного факта;

2) повышение точности тестирования знания отдельного факта, дающее возможность сократить количество тестовых заданий.

Работы в первом направлении по сокращению времени на всех трех шагах по тестированию знания отдельных фактов активно ведутся и становятся де-факто стандартной практикой при тестировании знаний. Так, генерацию вопросов проводят заблаговременно эксперты вручную в избыточном количестве, формируя банки вопросов [5; 6]. Это позволяет в момент тестирования быстро собрать текст из вопросов, отобранных из банка по заданным критериям. Очевидно, что заранее подготовленные вопросы ограничивают свободу образовательных траекторий студента. В новых системах компьютерного контроля знаний ставится задача, чтобы компьютерная программа по желанию преподавателя могла мгновенно и в любых количествах создавать вопросы для тестирования знания любого конкретного факта. В направлении автоматической интеллектуальной генерации вопросов особенно продвинулись отечественные разработчики [7, 8]. На европейских симпозиумах также сообщается о разработках систем семантического анализа текстов, которые помогут автоматически генерировать вопросы по заданному фрагменту текста [9; 10].

Оптимальное предъявление заданий студенту с учетом контроля временных затрат сводится к последовательному предъявлению отдельных заданий. Для минимизации влияния неучтенных факторов задания берут из банка вопросов в случайном порядке, исключая повторы [11]. Если задания разделены на несколько уровней сложности, то экономить время на тестирование позволяют разные алгоритмы адаптивного тестирования [3, 12].

Оптимизация получения ответа достигается предъявлением наряду с вопросом готовых вариантов ответов (задания закрытого типа), что позволяет не тратить студенту время на припоминание ответа, если он его знает, и при этом не дает преимуществ, если он его не знает (при правильно подобранных дистракторах [6] и формуле оценки, учитывающей угадывание [2]). При необходимости использовать прямые ответы студента (задания открытого типа) можно

автоматизировать проверку ответов средствами семантического анализа [13; 14] При ограничении времени на тестирование важно давать студенту возможность перейти к другому заданию, если текущее поставило его в тупик и вынуждает его напрасно терять много времени [15]. В ходе тестирования не следует давать обратную связь о правильности решения, чтобы не терять время на осмысление возможных ошибок. При тестировании разных групп студентов с использованием заданий из одного банка вопросов для профилактики коллективного списывания обратную связь о правильности решения отдельных заданий лучше отложить до окончания тестирования в последней группе.

Среди разработок во втором направлении по повышению точности тестирования практическое распространение получило введение уточняющего вопроса про уверенность (Certainty-Based Marking (CBM)) [16]. Модификацией этого способа можно считать способ исключения неправильных вариантов (EA) [17; 18], который по сути сводится к варианту теста с возможностью выбора нескольких правильных ответов, когда напротив каждого подходящего варианта нужно поставить галочку. В данном случае для каждого варианта требуется выбрать оценку правдоподобия из набора «правильно», «не знаю», «неправильно». То есть данный подход является комбинацией теста с выбором нескольких правильных ответов и теста с дополнительной оценкой уверенности.

Упомянутые способы повышению точности тестирования в общем случае приводят к повышению затрат времени и нервного напряжения учащихся, особенно, если информация о набранных баллах предоставляется сразу после выбора ответа [19].

Также предпринимались попытки повышения точности оценки, основанные на измерении времени, затраченного на выполнение тестового задания [20]. Индекс усилий для прохождения теста (Test-Taking Effort Index (TEI)) рассчитывается как разница во времени неправильных и правильных ответов [21]. Способ основан на предположении, что на неправильный ответ добросовестный студент тратит больше времени (сомневается, вспоминает), чем на правильный. Однако, если тестовое задание требует процедурного мышления, то время решения будет пропорционально количеству требуемых мыслительных операций, то есть процедурной сложности задания. Также наблюдается обратная зависимость при стратегии случайного угадывания: много неправильных ответов при минимальных затратах времени.

Таким образом, в большинстве известных методов повышения точности оценки знаний анализируется информация дополнительных показателей, на получение которых требуется дополнительное время (напр. оценка уверенности). Попытка использовать время, затраченное на тестовое задание, показала свою перспективность, но только в ограниченных условиях. Простой информации о затраченном времени оказалось недостаточно для оценки при разных стратегиях.

Ответ на вопрос теста множественного выбора – многокомпонентная деятельность, состоящая как минимум из просмотра вопроса, просмотров вариантов ответа, отметки правильного ответа. При неуверенности в ответе просмотры элементов тестового задания могут быть многократно повторены, а отмеченный вариант ответа несколько раз заменен на другие. В литературе отсутствуют примеры использования информации о паттерне выполнения тестового задания, хотя в результате анализа составляющих его элементарных операций можно было бы извлечь дополнительную информацию для повышения точности оценки знаний.

В данной работе была поставлена задача: разработать метод повышения эффективности контроля знаний с использованием информации о динамике поведения студента в процессе тестирования.

Методология и реализация

Для репрезентативности тестовой выборки (включения случаев решения тестовых заданий с различными паттернами поведения студентов) были проанализированы данные, набранные при пересдаче студентами 4 курса контрольных тестов по предмету «Физиология человека и животных». Расчет был такой, что, поскольку студенты не смогли сдать тест с первого раза, то и на пересдаче многие из них будут с низким уровнем знаний, что подвигнет их прибегнуть к разнообразным стратегиям прохождения теста.

У студентов на руках были учебные материалы с перечнем всех вопросов, поэтому при желании они могли ознакомиться со всеми вопросами заранее. Все участники были хорошо знакомы с процедурой тестирования в рамках электронной обучающей системы и дали информированное согласие на участие в нем.

Тестирование проводилось с помощью программного обеспечения Moodle для организации электронного обучения [23]. Формирование вопросов и проведение процедуры тестирования проводилось в соответствии с настройками по умолчанию. Тестовые задания предъявлялись по одному на странице. Для перехода к следующему заданию студенты нажимали клавишу «Далее», при этом они могли ориентироваться по панели навигации в том, сколько вопросов они уже прошли и сколько еще предстоит. По окончанию теста они могли просмотреть количество набранных баллов и содержание вопросов с отмеченными правильными и неправильными ответами.

Были собраны данные по тестированию 48 студентов по 12 темам: «ВНД», «Возбудимые ткани», «Дыхание», «Кровь», «Мышцы», «Нервная ткань», «Общая физиология ЦНС», «Основы регуляции», «Пищеварение», «Сенсорные системы», «Сердце, сосуды», «Частная физиология ЦНС». Некоторые студенты по некоторым темам пересдавали тест несколько раз. В общей сложности была собрана информация о просмотре 6610 тестовых заданий, из них 93.4% решенных, 6.6% - пропуски. В среднем по 128.7 решенных заданий (без учета пропущенных) на одного студента.

Итоговые оценки прохождения тестов для пересдачи были преимущественно положительные. Например, медиана оценок за тест «Возбудимые ткани (для пересдачи)» по 100-балльной шкале составила 85, средняя величина - 81,59, стандартное отклонение - 10,06 баллов. По другим тестам наблюдалась похожая статистика.

Важными характеристиками поведения студента в ходе единичного эпизода тестирования являются порядок и длительность составляющих элементарных операций. Так как фокус внимания человека ограничен, то в единицу времени происходит восприятие ограниченного кванта информации. Основными квантами информации при традиционном тестировании служат текст вопроса, который может включать дополнительные изображения, и текст предлагаемых вариантов ответа. Иерархия анализируемых элементарных операций показана на рис.1.

Нами предложен метод изоляции этих элементарных операций, когда текст вопросов и вариантов ответов закрывается непрозрачными экранами, которые становятся прозрачными при наведении курсора мыши. Таким образом, в один момент времени возможен просмотр только одного текстового элемента [22]. Методика позволяет измерять время просмотра каждого варианта ответа и паттерн перемещения между ними. Схему работы приложения см. на рис.2.

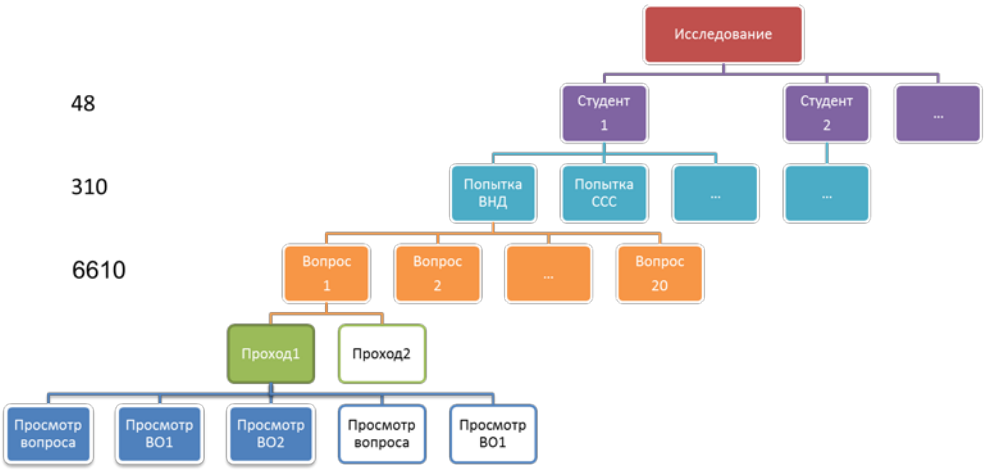

Рис. 1. Схема сбора данных для анализа элементарных операций
Цифры слева показывают количество экземпляров на соответствующем уровне иерархии в общей выборке данных. Фигуры без заливки - необязательные элементарные операции.

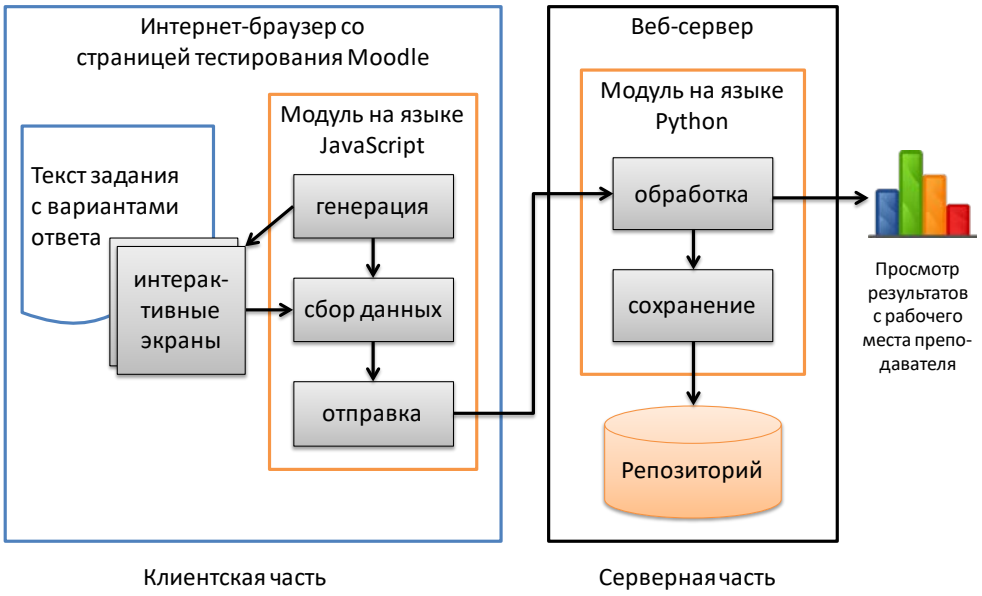

Рис. 2. Схема приложения для анализа элементарных операций

Время просмотра текстового элемента измерялось от момента наведения курсора, когда текст можно прочитать, до момента перевода курсора за пределы текстового элемента, когда экран вновь становится непрозрачным. В процессе тестирования студент целенаправленно перемещает курсор мыши к тому текстовому элементу, который хочет прочитать. Обычно при прохождении теста студент читает вопрос, затем переходит к чтению вариантов ответа. Если прочитанный им вариант ответа кажется ему правдоподобным, то он отмечает его как правильный, для чего кликает по нему мышкой (нажимает на кнопку мыши в момент нахождения курсора над текстом варианта ответа). Если в задании можно выбрать только один ответ, то после выбора правильного ответа оператор может не читать оставшиеся варианты и

закончить выполнение тестового задания, нажав на кнопку «Далее». При этом он переходит к следующему тестовому заданию или на страницу просмотра результатов. Если можно выбрать несколько вариантов ответов, то оператор просматривает каждый из вариантов ответа, и кликает мышкой на нем, если считает его правильным. Если студент сомневается в сделанном выборе, то он повторно просматривает вызвавшие сомнения варианты ответа. Также он может повторно прочитать текст вопроса. Если студент не точно помнит то, о чем спрашивается в вопросе, то он для полноты картины быстро знакомится с содержанием всех вариантов ответа, просматривая их подряд, не выбирая ни один из них (так называемое сканирование). Наконец, если студент не обладает знаниями, необходимыми для ответа на вопрос, то он может кликнуть мышкой наугад на любом из вариантов ответа. Таким образом, стратегия, выбранная оператором для прохождения данного тестового задания, отражается в порядке и скорости следования событий просмотра текстовых элементов и нажатий на кнопку мыши. Примеры с демонстрацией развертки во времени при различных стратегиях см. в нашей предыдущей работе с описанием метода [22].

Обработка данных с временными отметками событий заключалась в классификации решенных тестовых задания на классы:

– уверенное знание Q_K;
– неуверенное знание Q_I;
– частичное знание, попытка догадаться путем сопоставления вопроса и вариантов ответа Q_P;
– случайное угадывание Q_G;
– списывание Q_C.

Многоступенчатый анализ временных маркеров событий позволил выделить 19 показателей выполнения тестового задания, характеризующих последовательность и скорость просмотра вопроса и вариантов ответа. На первой ступени выделялись паттерны перемещений по текстовым элементам, такие как сканирование (быстрый последовательный просмотр всех вариантов ответа без принятия решения). На второй ступени рассчитывались интегральные показатели, такие как средняя скорость чтения по всем операциям чтения, количество пауз, а также суммарные показатели, такие как общее количество букв по всех тестовых элементах, длительность проверки (включает все операции после окончательного клика мыши). Ряд показателей были введены для контроля влияния внешних факторов: идентификационный код студента, доля суток для учета циркадианных воздействий, порядковый номер тестового задания, оценка.

В разработанном прототипе системы для классификации на пять стратегий прохождения теста методом кросс-валидации были отобраны 7 наиболее информативных показателей. 10-кратная кросс-валидация проводилась методом случайного леса (Random Forest) при обучении 300 случайных деревьев (estimators) на частичных обучающих выборках из решенных заданий, размеченных экспертом по принадлежности к пяти стратегиям. По мере уменьшения информативности показатели представлены в таблице 1.

Классификацию стратегий с использованием отобранных интегральных показателей можно реализовать с помощью несложных правил.

Например, уверенное знание Q_K характеризуется положительной оценкой $grade = 1$ и отсутствием избыточных просмотров вариантов ответов $na_{extra} \approx 0$. При этом коэффициент пропорциональности усилий k_{effort} должен принимать некоторое стандартное значение: затраты времени для принятия решения должны быть достаточны для чтения и осмысления прочитанного текста.

Таблица 1.

Интегральные показатели выполнения тестового задания

Название	Код	Информативность, у.е.
Доля бездействия (времени, не занятого просмотром текста)	p_{idle}	0.23
Время первого клика мыши, с	tc_{first}	0.225
Коэффициент пропорциональности усилий, равен отношению времени на принятие окончательного решения (время последнего клика мыши) к объему текста во всех текстовых элементах.	k_{effort}	0.205
Минимальная скорость просмотра текстового элемента, симв/с	v_{min}	0.19
Количество избыточных просмотров вариантов ответов.	na_{extra}	0.072
Формальная оценка в диапазоне 0...1, импортируется из системы обучения. Может быть дробной при выборе не всех правильных ответов, если их несколько.	grade	0.071
Доля времени, проведенная вне окна с тестом. Больше нуля, если оператор переключался в другое окно или другую вкладку интернет-браузера.	p_{out}	0.011

В случае, если студенту посчастливилось попасть в правильный вариант ответа $grade = 1$ методом случайного угадывания Q_G без просмотра других вариантов ответа, то это легко дифференцировать по отрицательному отклонению в количестве просмотров вариантов ответов $na_{extra} < 0$. Если студент спешит, например, когда время, отведенное на тест, подходит к концу, и даже не пытается прочитать текст вопроса, но выбирает наудачу какой-нибудь вариант ответа, то такое «решение» дифференцируется по очень низкому коэффициенту пропорциональности усилий k_{effort}.

Положительное отклонение в количестве просмотров вариантов ответов $na_{extra} > 0$ свидетельствует о наличии остаточных частичных знаний у студента Q_P, с помощью которых он пытается сопоставить варианты ответов друг с другом и с содержанием вопроса. Как правило это приводит к отбраковке заведомо неправильных вариантов ответа и сосредоточению на двух наиболее правдоподобных вариантах, тем самым повышая вероятность угадывания до 50%. Дополнительные просмотры вариантов ответов $na_{extra} > 0$ уже после выбора правильного ответа соответствуют проверке правильности своего решения, когда из-

за неуверенного знания Q_I требуется еще раз сопоставить варианты ответа. О наличии знаний при этом свидетельствуют положительная оценка $grade = 1$ и повышенный коэффициент пропорциональности усилий k_{effort}.

Однако для быстрого нахождения правил, оптимально разделяющих классы, предпочтительнее использовать методы машинного обучения, например, обучения классификатора методом случайного леса из деревьев решений.

Для апробации метода мы обучали классификатор на размеченной экспертом выборке методом стохастического градиентного бустинга, реализованным в пакете scikit-learn [24]. Выборку в 500 решенных тестовых заданий подбирали так, чтобы в ней были представлены все пять стратегий выполнения заданий, при этом другие факторы рандомизировались (рис.3). Также в выборку не вошли случаи с паузами более 15 с. В обучающую выборку вошли в среднем по 10.6 решений 47 разных студентов. Метод стохастического градиентного бустинга позволяет быстро обучить классификатор так, что каждое добавляемое дерево решений (estimator) улучшает решение, найденное при обучении предыдущих. При этом не требуется преобразование показателей к одной размерности, что актуально в нашем случае. При обучении использовались параметры, обеспечивающие 100% аккуратность классификации при обучении: максимальная глубина – 3, минимальная расщепляемая выборка – 5, общее количество деревьев – 300. Средняя ROC AUC полученная при этих параметрах при кросс-валидации по 5 сложениям составила для каждого из пяти классов против остальных, соответственно C - 0.92, G - 0.93, I - 0.88, K - 0.86, P - 0.88. Общая аккуратность классификатора составила 0.71. Классификатор на выходе дает значения вероятностей принадлежности $[p_C \quad p_G \quad p_I \quad p_K \quad p_P]$ к пяти классам стратегий выполнения тестовых заданий, соответственно: списывание Q_C, случайное угадывание Q_G, неуверенное знание Q_I, уверенное знание Q_K, частичное знание Q_P.

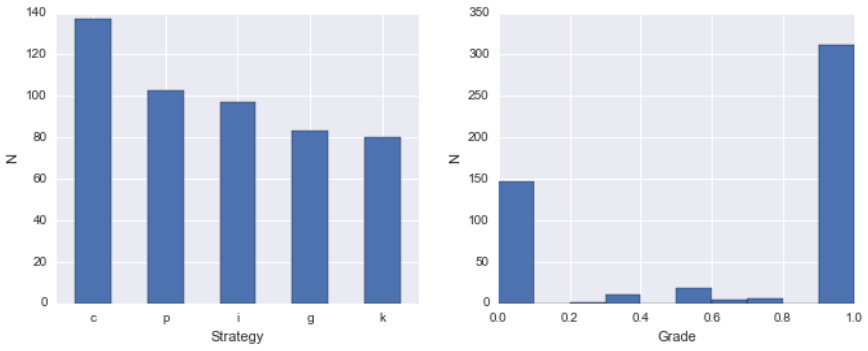

Рис. 3. Распределение показателей в обучающей выборке
А – выделенные экспертом стратегии, Б – оценки

Информация о стратегии, выбранной студентом при прохождении теста, полезна для преподавателя. При этом информация должна быть представлена в наглядном виде. Стратегией выполнения одного тестового задания можно считать класс с наибольшей вероятностью принадлежности, а преобладающей стратегией выполнения всего теста – стратегию, выявленную при выполнении большинства тестовых заданий в данном тесте. Для наглядности распределение выявленных стратегий при выполнении заданий в рамках одного теста может быть представлено в виде круговой диаграммы.

Массив значений вероятностей принадлежности можно интегрировать в оценку, установив балльное соответствие каждой из стратегий. После этого расчет итоговой оценки G_i можно провести по формуле центра масс:

$$G_i = \frac{\sum_{i=K,I,P,G,C} g_i p_i}{\sum_{i=K,I,P,G,C} p_i}$$

где p_i - вероятность принадлежности к i-той стратегии выполнения тестовых заданий, g_i – оценка за выполнение тестового задания по i-той стратегии: $g_K = 1.0, g_I = 0.67, g_P = 0.33, g_G = g_C = 0$.

Предлагаемые коэффициенты для расчета интегральной оценки в общем соответствуют стандартной схеме оценивания, когда любые действия пользователя оцениваются в диапазоне [0…1]. При этом одинаково высокую оценку 1 студент получает за знание ответа, удачное угадывание или списывание. В предлагаемой нами системе оценки будут систематически ниже стандартных за счет выявления и исключения необоснованно высоких оценок при угадывании правильных ответов, списывании и неуверенности. Порог для бинарной оценки зачет/незачет для такой шкалы целесообразно задать около 0.5, что ниже общепринятого в России порога 0.6.

Анализ и оценка разработки

Метод был апробирован при оценке знаний студентов по дисциплине «Физиология человека и животных» на кафедре физиологии Академии биологии и биотехнологии Южного федерального университета.

Для этого мы провели расчеты оценки знаний за отдельные попытки пересдачи контрольных тестов с учетом системы анализа поведения студента и без нее.

Все тесты по отдельным темам состояли из 20 вопросов. При каждой попытке пройти тест проводился случайный отбор тестовых вопросов из банка по около 100 вопросов по каждой теме. Тестовые вопросы не разделялись по сложности и при расчете оценок имели равные весовые коэффициенты. Длительность прохождения одного теста была ограничена 20 минутами.

Для удобства академические оценки за ответы в тестах приводятся в 100-балльной шкале. Итоговая оценка за тест рассчитывалась стандартно как среднее значение двадцати оценок. За невыполненные задания давалось 0 баллов.

Для сравнения мы выбирали конкретную попытку пройти тест и сопоставляли результаты стандартной системы оценивания, встроенной в обучающую систему, и результаты авторской системы оценивания, описанной выше. Для этого мы проводили подготовку показателей для каждого из 20 вопросов данной попытки, причем ознакомительные проходы без принятия решений (если такие были) не учитывались. Далее с помощью обученного классификатора мы предсказывали класс стратегии при решении каждого из заданий, а также рассчитывали интегральную оценку. В качестве итоговой оценки представлено среднее значение двадцати интегральных оценок, приведенное к 100-балльной шкале.

В качестве первого примера рассмотрим результаты студента, который хорошо подготовился к пересдаче по теме «ВНД» (высшая нервная деятельность)(рис.4). Он отвечал на 20 вопросов последовательно без пропусков и, затратив 9 мин 49 с, набрал 93 балла. Система оценки с выделением стратегий отнесла по половине выполненных заданий к классам уверенного и неуверенного знания. В итоге получилась высокая оценка, свидетельствующая о наличии знаний у данного студента по данной теме.

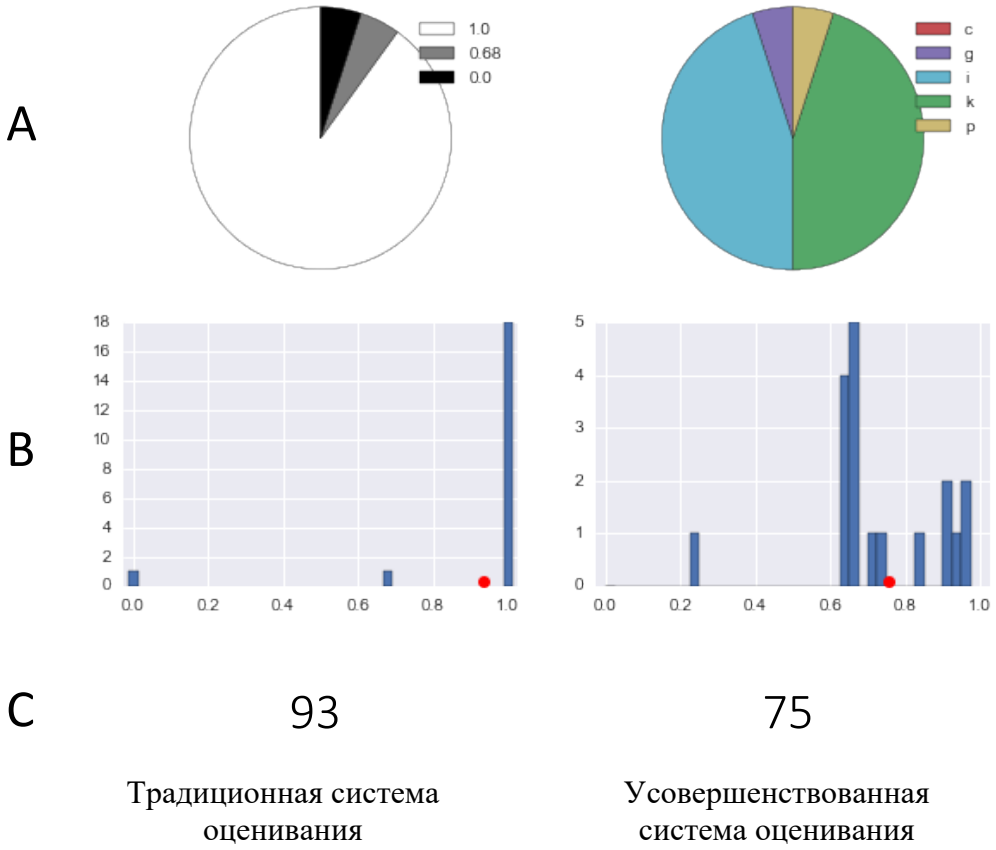

Рис. 4. Результаты тестирования хорошо подготовленного студента
А – соотношение выделяемых классов выполнения тестовых заданий; В – распределение оценок, по оси абсцисс – значение, по оси ординат – количество. Красная точка – средневзвешенное значение. С – оценка по 100-балльной шкале. Слева – традиционная система оценивания, справа – авторская.

В качестве другого примера рассмотрим вариант плохого выполнения теста. Студент выполнял задание в течение 9 мин. 44 с по теме «Сенсорные системы». Несмотря на то, что у него еще было время подумать до 20 мин, он закончил тест раньше, дав в основном неправильные ответы. Стандартная система оценки посчитала 5 правильных ответов и 2 частично правильных с итоговой оценкой в 30 баллов.

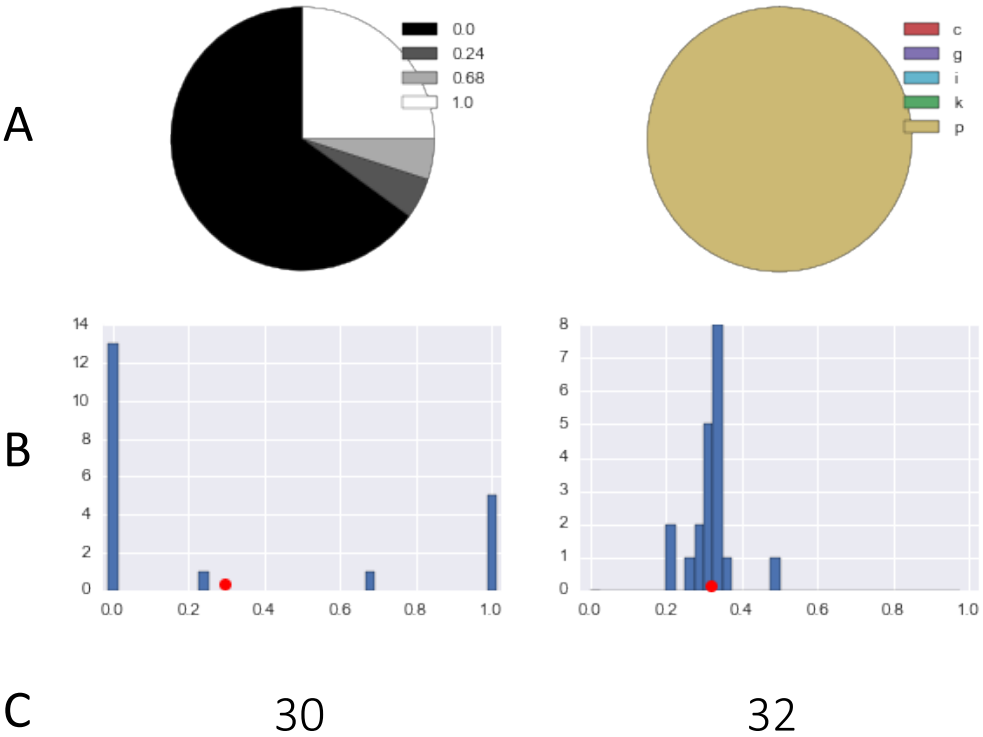

Рис. 5. Результаты тестирования плохо подготовленного студента
А – соотношение выделяемых классов выполнения тестовых заданий; В – распределение оценок, по оси абсцисс – значение, по оси ординат – количество. Красная точка – средневзвешенное значение. С – оценка по 100-балльной шкале. Слева – традиционная система оценивания, справа – авторская.

Усовершенствованная система оценки с выделением стратегий дает удивительно близкую оценку – 32 балла. Однако распределение индивидуальных оценок (рис. 5Б справа) наглядно показывает, что студент в ходе всего теста придерживался одной выбранной стратегии – угадывание с опорой на остаточные знания. Т.е. он читал вопросы и просматривал варианты ответов, пытаясь выполнить тест, но из-за отсутствия знаний почти все ответы были неправильны.

Следующий пример призван продемонстрировать случай, когда стандартная и усовершенствованная системы оценивания дают принципиально разные результаты. На рисунке 6 представлены оценки одного из студентов за пересдачу по теме «Сенсорные системы». Данный студент видимо очень спешил - весь тест занял всего 7 мин 9 с. При этом он получил положительную оценку в 60 баллов.

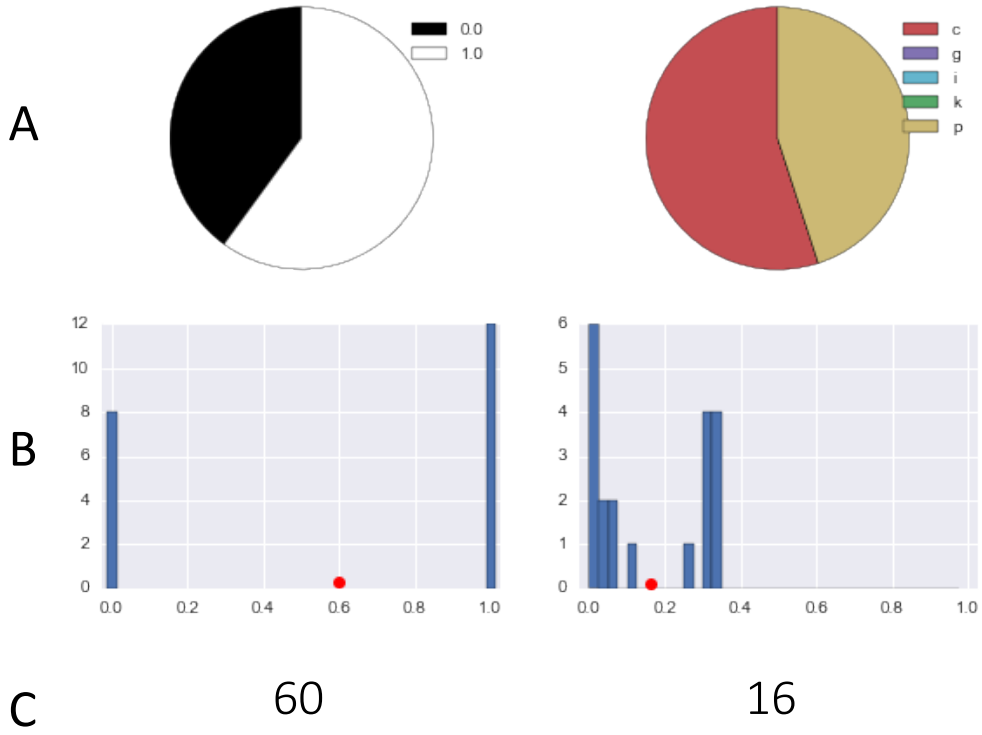

Рис. 6. Результаты тестирования со смешанной стратегией выполнения теста
A – соотношение выделяемых классов выполнения тестовых заданий; B – распределение оценок, по оси абсцисс – значение, по оси ординат – количество. Красная точка – средневзвешенное значение. C – оценка по 100-балльной шкале. Слева – традиционная система оценивания, справа – авторская.

В данном случае студент для нахождения правильных ответов пользовался внешними источниками. Паттерн выполнения заданий при этом состоял из долгого просмотра вопроса 13-41 с, а затем из очень быстрого просмотра вариантов ответа с выбором правильных. Решение заданий с таким паттерном были распознаны классификатором как относящиеся к классу C – списывание (cheating).

Если стандартная система оценивания за правильный ответ дает 1 балл, независимо от того, дал его студент сам или позаимствовал из внешнего источника, то используемая в данной работе авторская система оценивания дает за списывание 0 баллов. На гистограмме интегральных оценок (рис. 6B справа) виден пик со значениями около нуля, относящихся к списыванию, однако на распределении заметен другой пик, относящийся к решениям заданий, распознанных как угадывание с опорой на частичные знания, класс P (partial). Так как при списывании ответы

формально правильные, то 40% неправильных ответов, которые показывает стандартная система оценивания (рис. 6А слева), относились к угадыванию.

На рисунке 7 представлены результаты этого же теста в динамике.

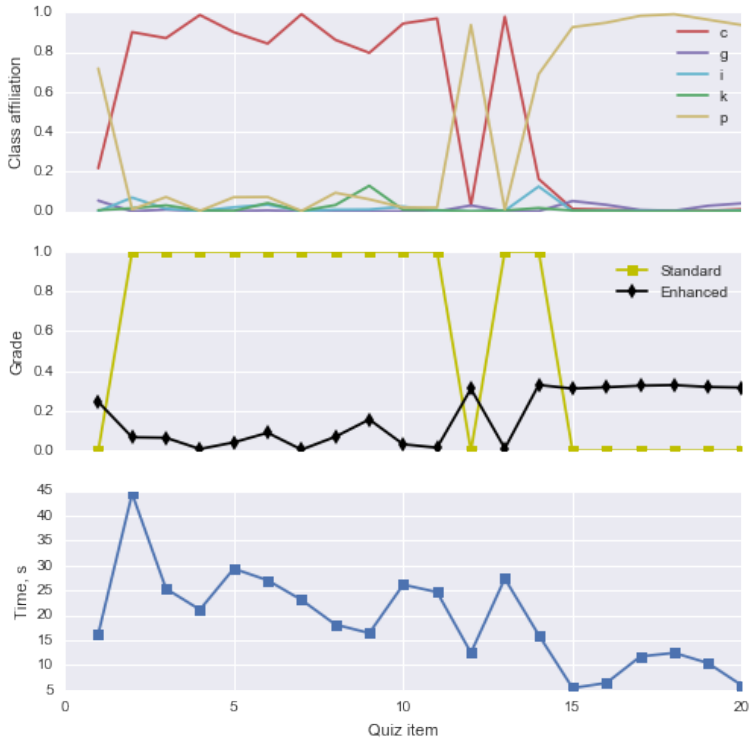

Рис. 7. Пример смены стратегий решения тестовых заданий в динамике теста
Вверху – оценка принадлежности к каждому из пяти классов стратегий: c - списывание, g - случайное угадывание, i - неуверенное знание, k - уверенное знание, p - частичное знание. В середине – оценки за тестовые задания: темная линия с ромбами – рассчитанные усовершенствованной системой оценивания, светлая линия - выставленные стандартной системой оценивания. Внизу – длительность выполнения тестового задания, с. По оси абсцисс – порядковый номер тестового задания.

В первой половине теста решение заданий стабильно классифицировалось как списывание, кроме 1-го и 12-го вопросов, когда паттерн решения был похож на угадывание. После долгого поиска ответа на 13-й вопрос, что заняло 28 с, студент попытался ответить сам. Паттерн ответа на 14 вопрос резко отличается отсутствием длительного просмотра текста вопроса (рис. 8). На хронограмме наглядно видно, как разворачивались события по ходу времени. Многократный просмотр вариантов ответов с поиском наиболее правдоподобного – типичный признак стратегии угадывания с опорой на остаточные знания.

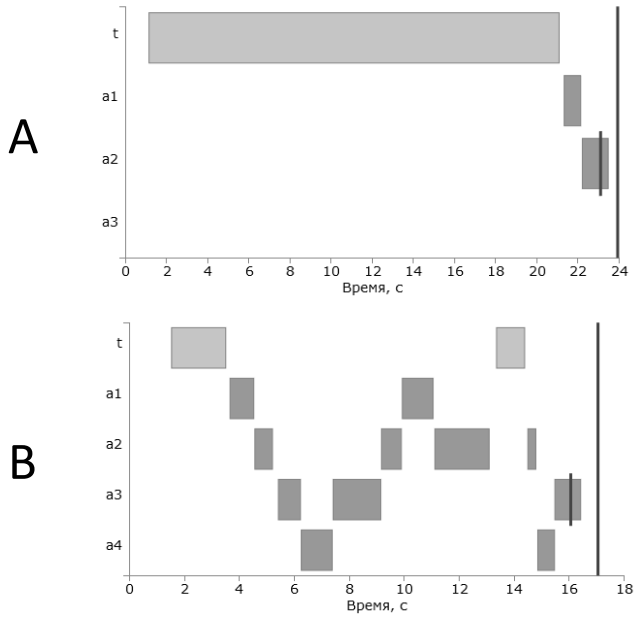

Рис. 8. Хронограмма решений до и после смены стратегии

A – вопрос №7, стратегия списывания, правильный ответ; B - вопрос №14, стратегия угадывания с опорой на частичные знания, правильный ответ. t –просмотр текста вопроса, aN – просмотр N-ного варианта ответа. По оси абсцисс – время от начала предъявления тестового задания, с. Вертикальной чертой показан момент клика мышкой. Большая вертикальная линия – конец выполнения тестового задания.

После правильного ответа на 14-й вопрос, когда студент набрал необходимые 12 правильных ответов, что соответствует 60 баллам, он решил сменить стратегию. Вместо долгого поиска ответа во внешнем источнике, он переключился на быстрое угадывание. Начиная с 15-го вопроса время ответа на вопрос не превышало 12.5 с (рис.7, внизу). Студент попытался ответить на оставшиеся вопросы сам, однако это ему не удалось: ответы на шесть последних вопросов теста были неверными.

Наибольшее расхождение между стандартной и усовершенствованной системами оценивания наблюдалось в случае 100% списывания. В этом случае официальная оценка на образовательном портале приближалась к 100 баллам, а оценка, рассчитанная с учетом выявления стратегий по предложенной формуле, приближалась к 0.

Рассмотренные примеры сравнения стандартной системы оценивания и усовершенствованной системы оценивания показывают, что метод, предложенный авторами, не противоречит стандартной практике тестирования, а дополняет ее. В случае добросовестного выполнения тестовых заданий оценки оказываются близки, однако, в случаях различных злоупотреблений со стороны тестируемого предлагаемый метод дает более адекватную оценку.

Новый метод основан на анализе информации о последовательности элементарных операций, осуществляемых в ходе тестирования. Авторская методика хронометрии элементарных операций позволяет регистрировать моменты начала и окончания просмотров текстовых элементов тестового задания множественного выбора. Эта информация является исходной для дальнейшего многоступенчатого анализа данных, заканчивающегося определением одной из пяти стратегий прохождения тестов. Таким образом, сфера потенциального применения

предложенного метода охватывает все случаи тестирования знаний, где уместны тестовые задания множественного выбора. Задания закрытого типа широко используются в практике высшего образования, профориентации, квалификационной аттестации.

В текущей реализации метода определение стратегий выполнения тестовых заданий производится с помощью специально обученного классификатора. Авторы выбрали тип классификатора, подобрали к нему параметры, отобрали наиболее информативные показатели, оценили работоспособность алгоритма на реальных данных. Однако, это не исключает возможности дальнейшего усовершенствования подхода к определению стратегий. Для других областей применения оптимальный набор информативных показателей может оказаться другим. Набор стратегий также может быть пересмотрен для более адекватной подстройки к требованиям других дисциплин.

Пока рано говорить о возможности универсального применения предложенного метода в том виде, в каком он представлен в данной работе, для повсеместного повышения эффективности тестирования знаний студентов. Метод хорошо подойдет для тестирования знаний понятийного аппарата, закономерностей, умения производить быстрые операции в уме, то есть для таких тестов, где требуются быстрые ответы без привлечения внешних приложений. Для других областей тестирования, например, где требуется решать расчетно-графические задачи, метод потребует существенной доработки, поскольку в тех условиях не будут соблюдаться изученные на данный момент закономерности по соответствию паттернов выполнения заданий и стратегий.

В данной работе сделана попытка показать, что может дать предложенный метод педагогу. Предложена формула расчета более точной оценки, интегрирующая вероятности принадлежности данного эпизода тестирования к каждой из выделенных стратегий. Простая оценка в баллах комфортнее для восприятия, чем указание перечня использованных стратегий или просмотр паттернов хронограмм. Сопоставление на конкретных примерах показывает явное преимущества описанной в работе усовершенствованной системы оценивания над стандартной общепринятой системой оценивания. Однако, можно предложить множество других формул расчета интегральной оценки в зависимости от задач, стоящих перед организаторами тестирования. Например, при строгом рубежном контроле выявленные факты списывания могут караться отрицательными баллами, и, наоборот, в тестировании текущего контроля для самопроверки поиск информации в Интернет может оцениваться более высоким баллом, чем попытка угадать с опорой на остаточные знания.

Информация о стратегиях выполнения тестовых заданий может быть интересна сама по себе. Предложенный метод может дать информацию для объяснения фактов резкого изменения в успеваемости. Вкупе с показателями расхода времени на тестирование, это может дать информацию о количестве усилий, затраченных студентом на учебу. Описанные интегральные показатели могут использоваться для оценки динамики развития навыков участия в тестировании (test-taking skill). Иными словами, дальнейшее развитие аналитического аппарата в этом направлении может дать начало новым методикам усовершенствования учебного процесса. Примечательно, что для получения столь ценной дополнительной информации достаточно небольшой модификации тестовой среды без привлечения дорогостоящего оборудования.

Имея в своем распоряжению информацию о стратегиях прохождения теста, педагог может сам выбрать желаемую форму представления этой информации. Если он доверят алгоритмам машинного обучения, если сам участвовал в экспертной оценке обучающей выборки, то он может ввести в практику уточненные оценки в диапазоне 0…1. И, наоборот, если он придерживается стандартных проверенных

временем подходов, то может оставить систему оценок без изменений, а автоматизированную систему определения стратегий выполнения тестовых заданий использовать для генерации предупреждений о вероятности нарушений. Например, при обнаружении нехарактерных паттернов в поведении студента во время тестирования, может подаваться сигнал преподавателю, после которого он может прибегнуть к традиционным контрольным мерам: просмотреть видеозапись из дисплейного класса, или перепроверить знания студента с помощью дополнительных заданий.

Выделение стратегий решения каждого задания позволяет интерпретировать поведение студентов в динамике. Как показано в последнем примере, в ряде случаев возможна смена стратегии в ходе теста. Особенно типична смена стратегии при приближении срока принудительного окончания тестирования в условиях цейтнота. Анализ динамики смены стратегий может дать ценную обратную связь организаторам образовательного процесса о неуспевающих студентах.

Неспособность справиться с тестом может быть вызвана не незнанием предмета, а другим неучтенным фактором, например, медленным чтением с экрана монитора у слабовидящего учащегося, или плохим знанием языка у иностранного студента. Предложенная в данной работе методика анализа поведения студентов в ходе тестирования может использоваться для сертификации тестовых материалов для особых категорий учащихся.

Повышенная вариация стратегий возможна, если в тесте встречаются задания различной сложности. В этом случае с помощью предложенной методики можно различить, какие задания даются студенту легко, какие не очень, а какие он даже не пытается выполнять, переходя к стратегии угадывания. Таким образом, открываются перспективы тонкой калибровки подготовленных банков тестовых заданий по сложности на основании относительно небольшого количества предварительных испытаний.

Картина смешения стратегий уверенных ответов и угадывания может также наблюдаться, если студент изучил только часть требуемого учебного материала. В этом случае при объективной одинаковой сложности заданий в восприятии студента эти две группы заданий будут субъективно различаться по сложности, склоняя его к разным стратегиям прохождения теста.

Заключение

В работе представлен метод повышения точности тестирования знаний учащихся с использованием заданий множественного выбора. Метод основан на автоматизированном выделении стратегий прохождения теста за счет анализа паттерна выполнения элементарных операций в ходе тестирования. Метод является перспективным для внедрения в образовательный процесс для объективного выявления стратегий учащихся.

Информация о стратегиях, выбранных студентом при решении заданий, может быть интегрирована в оценку по успеваемости, или носить рекомендательный характер для педагога, отслеживающего динамику развития компетенций студента. Интеграция в оценку по успеваемости значительно повышает точность тестирования, а значит, позволяет сократить время на текущий контроль развития профессиональных компетенций. Это очень важно в условиях интенсификации современного высшего образования.

Работа выполнена в рамках ГЗ по теме №213.01-11/2014-35.

Литература

1. Галеев И.Х. Компьютерный контроль знаний (локально и дистанционно) / И.Х. Галеев, В.Г. Иванов, Д.Л. Храмов, О.В. Колосов; Под ред. И.Х. Галеева. - Казань: Казанский государственный технологический университет, 2005. – 126с.
2. Espinosa M. P., Gardeazabal J. Optimal correction for guessing in multiple-choice tests //Journal of Mathematical Psychology. – 2010. – Т. 54. – №. 5. – С. 415-425.
3. Галеев И.Х., Храмов Д.Л., Светлаков А.П., Колосов О.В. Адаптивное обучение и тестирование. //Материалы Всероссийской научно-методической конференции «Развитие методов и средств компьютерного адаптивного тестирования», 17-18 апреля 2003 г. – С. 33-35.
4. Weiss, D.J., ed. New Horizon Testing: Latent Trait Test Theory and Computerized Adaptive Testing. Elsevier, 2014.
5. Burton R. F. Multiple-choice and true/false tests: myths and misapprehensions //Assessment & Evaluation in Higher Education. – 2005. – Т. 30. – №. 1. – С. 65-72.
6. Bush M. E. Quality assurance of multiple-choice tests //Quality Assurance in Education. – 2006. – Т. 14. – №. 4. – С. 398-404.
7. Комплекс программ для компьютерного тестирования http://www.ast-centre.ru/testirovanie/ast_test/ (дата обращения: 05.08.2016)
8. Галеев И.Х., Иванов В.Г., Аристова Н.В., Урядов В.Г. Сравнительный анализ программных комплексов TestMaker и ACT-Test // Международный электронный журнал "Образовательные технологии и общество (Educational Technology & Society)" - 2007 - V. 10 -N 3. - С.336-360. - ISSN 1436-4522. URL: http://ifets.ieee.org/russian/periodical/journal.html
9. Teo N. H. I., Bakar N. A., Karim S. Designing GA-Based Auto-Generator of Examination Questions //Computer Modeling and Simulation (EMS), 2012 Sixth UKSim/AMSS European Symposium on. – IEEE, 2012. – С. 60-64.
10. Патент US2015199400 15.01.2014 Wu P. Automatic generation of verification questions to verify whether a user has read a document / заяв. пат. 14/156,152 США. – 2014.
11. Переверзев В.Ю. Критериально-ориентированное педагогическое тестирование в профессиональном образовании (методология, теория, практика). М. ФИРО, 2008. 247 с.
12. Truong H. M. Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities //Computers in Human Behavior. – 2016. – Т. 55. – С. 1185-1193.
13. Kardan A. A., Sani M. F., Modaberi S. Implicit learner assessment based on semantic relevance of tags //Computers in Human Behavior. – 2016. – Т. 55. – С. 743-749.
14. Sultan M. A., Salazar C., Sumner T. Fast and Easy Short Answer Grading with High Accuracy //Proceedings of NAACL-HLT. – 2016. – С. 1070-1075.
15. Wang T., Zhang J. Optimal partitioning of testing time: theoretical properties and practical implications //Psychometrika. – 2006. – Т. 71. – №. 1. – С. 105-120.
16. Bereby-Meyer Y., Meyer J., Flascher O. M. Prospect theory analysis of guessing in multiple choice tests //Journal of Behavioral Decision Making. – 2002. – Т. 15. – №. 4. – С. 313.
17. Lau P.N., Lau S.H., Hong K.S., Usop H. Guessing, partial knowledge, and misconceptions in multiple-choice tests //Journal of Educational Technology & Society 14, no. 4 – 2011. – С. 99-110. http://www.jstor.org/stable/jeductechsoci.14.4.99.
18. Chang S. H. et al. Measures of partial knowledge and unexpected responses in multiple-choice tests //Educational Technology & Society. – 2007. – Т. 10. – №. 4. – С. 95-109.
19. Schoendorfer N., Emmett D. Use of certainty-based marking in a second-year medical student cohort: a pilot study //Advances in medical education and practice. – 2012. – Т. 3. – С. 139.

20. Lee Y. H., Haberman S. J. Investigating Test-Taking Behaviors Using Timing and Process Data //International Journal of Testing. – 2015. – С. 1-28.
21. Chang S. R. et al. Development and application of detection indices for measuring guessing behaviors and test-taking effort in computerized adaptive testing //Educational and Psychological Measurement. – 2011. – Т. 71. – №. 3. – С. 437-459.
22. Щербина Д.Н. Стратегии прохождения тестов знаний, выявленные методом хронометрии просмотра вариантов ответа // Валеология. – 2015. – № 4. – С. 112-121.
23. Система управления обучением Moodle URL: https://moodle.org/ (дата обращения: 18.06.2016).
24. Scikit-learn: machine learning in Python URL: http://scikit-learn.org. (дата обращения: 05.08.2016)